

1 October 2020, Vol. 50, No. 10

Online archive: <http://www.sigda.org/publications/newsletter>

1. [SIGDA News](#)

From: Xiang Chen <shawn.xiang.chen@gmail.com>

2. [SIGDA Local Chapter News](#)

From: Yiran Chen <yiran.chen@duke.edu>

From: Tsung-Yi Ho <ho.tsungyi@gmail.com>

3. ["What is" Column](#)

Contributing author: Zheng Dong <dong@wayne.edu>

From: Wanli Chang <wanli.chang@york.ac.uk>

4. [Paper Submission Deadlines](#)

From: Xin Zhao <xzhao@us.ibm.com>

5. [Upcoming Conferences and Symposia](#)

From: Xin Zhao <xzhao@us.ibm.com>

6. [Researcher Spotlight](#)

From: Rajsaktish Sankaranarayanan <rsankara@ieee.org>

7. [SIGDA Partner Journal](#)

From: Matthew Morrison <matt.morrison@nd.edu>

From: Jeff Goeders <jgoeders@byu.edu>

8. [Technical Activities](#)

From: Ying Wang <wangying2009@ict.ac.cn>

9. [Notice to Authors](#)

Comments from the Editors

Dear ACM/SIGDA members,

We are excited to present to you the October e-newsletter. We encourage you to invite your students and colleagues to be a part of the SIGDA newsletter. The newsletter covers a wide range of information from the upcoming conferences and hot research topics to technical news and activities from our community. Get involved and contact us if you want to contribute an article or announcement.

The newsletter is evolving, let us know what you think.

Happy reading!

[Debjit Sinha](#), Keni Qiu, Editors-in-Chief, SIGDA E-News

To renew your ACM SIGDA membership, please visit <http://www.acm.org/renew> or call between the hours of 8:30am to 4:30pm EST at +1-212-626-0500 (Global), or 1-800-342-6626 (US and Canada). For any questions, contact acmhelp@acm.org

SIGDA E-News Editorial Board:

[Debjit Sinha](#), E-Newsletter co Editor-in-Chief

Keni Qiu, E-Newsletter co Editor-in-Chief

Xiang Chen, E-Newsletter Associate Editor for SIGDA News column

Yanzhi Wang, E-Newsletter Associate Editor for SIGDA Local chapter news column

Pingqiang Zhou, E-Newsletter Associate Editor for SIGDA Awards column

Wanli Chang, E-Newsletter Associate Editor for SIGDA What is column

Xun Jiao, E-Newsletter Associate Editor for SIGDA What is column

Jayita Das, E-Newsletter Associate Editor for SIGDA Funding opportunities column

[Qinru Qiu](#), E-Newsletter Associate Editor for SIGDA Live column

Yiyu Shi, E-Newsletter Associate Editor for SIGDA Live column

Rajsaktish Sankaranarayanan, E-Newsletter Associate Editor for SIGDA Researcher spotlight column

Xin Zhao, E-Newsletter Associate Editor for SIGDA Paper submission deadline column

Ying Wang, E-Newsletter Associate Editor for SIGDA Technical activities column

[Back to Contents](#)

SIGDA News

(1) "It's Official: Nvidia Buys Arm"

[\[https://www.eetimes.com/nvidias-arm-acquisition-reportedly-close-to-announcement...\]](https://www.eetimes.com/nvidias-arm-acquisition-reportedly-close-to-announcement...)

Nvidia announced that it has signed a definitive agreement to acquire Arm from SoftBank, in cash-and-stock transaction valued at \$40 billion.

(2) "Qualcomm Cloud AI 100 Promises Impressive Performance per Watt for Near-Edge AI"

[\[https://www.eetimes.com/qualcomm-cloud-ai-100-promises-impressive-performance-pe...\]](https://www.eetimes.com/qualcomm-cloud-ai-100-promises-impressive-performance-pe...)

Almost 18 months on from the initial announcement of the Cloud AI 100 AI accelerator, Qualcomm has released a few further details of the solution form factors this chip will be available in, and a few performance figures for those cards. The mobile silicon giant revealed that Cloud AI 100 final silicon is in production and will ship in the first half of 2021.

(3) "Verizon Taps Partners to Build 5G Private Networks"

[\[https://www.eetimes.com/verizon-taps-partners-to-build-5g-private-networks/\]](https://www.eetimes.com/verizon-taps-partners-to-build-5g-private-networks/)

Private networks are becoming one of the biggest growth sectors in the global telecommunications market, a trend confirmed by the results of the recently concluded auction for private licenses to 3.5 GHz airwaves. Verizon, Comcast and Dish Networks got most of the spectrum (and the headlines), but as significant was the number of small groups and non-traditional network operators who gained surprisingly large chunks of the vital resource. They will be using the spectrum to build their own 4G LTE and 5G capable networks.

(4) "Ericsson Buys Cradlepoint to Enter Enterprise 5G Market"

[\[https://www.eetimes.com/ericsson-buys-cradlepoint-to-enter-enterprise-5g-market/\]](https://www.eetimes.com/ericsson-buys-cradlepoint-to-enter-enterprise-5g-market/)

Ericsson's acquisition of Cradlepoint should silence the detractors who have been suggesting for some time that the Swedish group has been neglecting the enterprise sector for too long in favor of other 5G opportunities.

(5) "NVMe-oF Is Ready to Go the Distance"

[\[https://www.eetimes.com/nvme-of-is-ready-to-go-the-distance/\]](https://www.eetimes.com/nvme-of-is-ready-to-go-the-distance/)

As an extension of the somewhat mature non-volatile memory express (NVMe) protocol, the NVMe-oF specification uses NVMe to connect hosts to storage across a network fabric. NVMe-oF supports the transfer of data between a host computer and a solid-state storage device or system over a network. Using a NVMe message-based command, these transfers can be done via Ethernet, Fibre Channel (FC), or InfiniBand.

(6) "Spin Partners with ARM, Applied in MRAM Manufacturing"

[\[https://www.eetimes.com/spin-partners-with-arm-applied-in-mram-manufacturing/\]](https://www.eetimes.com/spin-partners-with-arm-applied-in-mram-manufacturing/)

Spin Memory has partnered with ARM and Applied Materials to start making what the Fremont, California-based startup expects to win broad adoption in military, automotive and medical equipment for their MRAM solutions.

(7) "Memory Technologies Confront Edge AI' s Diverse Challenges"

[\[https://www.eetimes.com/memory-technologies-confront-edge-ais-diverse-challenges...\]](https://www.eetimes.com/memory-technologies-confront-edge-ais-diverse-challenges...)

With the rise of AI at the edge comes a whole host of new requirements for memory systems. Can today' s memory technologies live up to the stringent demands of this challenging new application, and what do emerging memory technologies promise for edge AI in the long-term?

(8) "DARPA: Research Advances for Near-Zero-Power Sensors"

[\[https://www.eetimes.com/darpa-research-advances-for-near-zero-power-sensors/\]](https://www.eetimes.com/darpa-research-advances-for-near-zero-power-sensors/)

Some battlefield sensors that used to run out of power in months (if not weeks) can now keep providing valuable intelligence for up to four years before their coin batteries need to be replaced. That' s one of the more dramatic results of DARPA' s Near Zero Power RF and Sensor Operations (N-ZERO) program, which concluded in May 2020.

(9) "Silicon Labs Adds Secured Bluetooth Low Energy Modules"

[\[https://www.eetimes.com/silicon-labs-adds-secured-bluetooth-low-energy-modules/\]](https://www.eetimes.com/silicon-labs-adds-secured-bluetooth-low-energy-modules/)

Silicon Labs has formalized its new BGM220x module with a size of only 6×6 mm. BGM220 is an embedded solution that comes with a fully upgradeable software stack, which has been pre-certified worldwide, and firmware support to accelerate time-to-market.

(10) "First 4D Imaging Radar Sensors for ADAS to Ship in Vehicles in 2021"

[\[https://www.eetimes.com/first-4d-imaging-radar-sensors-for-ad-as-to-ship-in-vehic...\]](https://www.eetimes.com/first-4d-imaging-radar-sensors-for-ad-as-to-ship-in-vehic...)

Continental announced it is using Xilinx FPGAs to deploy the automotive industry' s first production-ready 4D imaging radar, expected to ship in passenger vehicles in 2021. Continental' s new advanced radar sensor (ARS) 540 will use the Zynq UltraScale+ MPSoC platform, enabling vehicles equipped with the sensor to realize SAE J3016 Level 2 functionalities, paving the way toward eventual Level 5 autonomous driving systems.

[Back to Contents](#)

SIGDA Local Chapter News

This Design Automation WebiNar (DAWN), the fourth event of the webinar series, will be on the topic of Publishing in EDA Transactions, Journals, and Magazines. It will consist of five excellent talks followed by Q&A with the panelists. We are thrilled to offer you the opportunity to attend DAWN and encourage your participation. The panel includes:

-- Rajesh Gupta - Editor-in-Chief of IEEE Transactions on Computer-Aided Design of Integrated

Circuits and Systems

- X. Sharon Hu - Editor-in-Chief of ACM Transactions on Design Automation of Electronic Systems
- Tulika Mitra - Editor-in-Chief of ACM Transactions on Embedded Computing Systems
- Ramesh Karri - Editor-in-Chief of ACM Journal of Emerging Computing Technologies
- Jörg Henkel - Editor-in-Chief of the IEEE Design&Test Magazine

Moderator:

Hai "Helen" Li, Professor, ECE Department, Duke University

Please visit <https://duke-cei-lab.github.io/DAWN/> for more information on the DAWN series.

[Back to Contents](#)

"What is" Column

What is Shared-Resource-Centric Real-Time Task Scheduling?

Dr. Zheng Dong
Assistant Professor
Department of Computer Science
Wayne State University, USA

In recent years, the microprocessor industry has turned to multi-core processor designs for the next generation of embedded systems. By increasing the number of cores, it is possible to dramatically improve the performance as well as energy efficiency. This trend has brought in the many-core architectures, which incorporate a large number of cores to provide unprecedented advantages in terms of performance-per-watt and inter-core communication latency over traditional microprocessor architectures. Examples of many-core platforms include Intel SCC [1], Godson-T [2], and STHorm [3], which have cores organized in multiple islands sharing various shared resources, such as DRAM modules, on-chip buses, or computing accelerators such as GPU and FPGA. Many-core platforms have been widely used in real-time embedded systems and proved to provide excellent performance for many applications such as computer vision and image analysis workloads.

To guarantee real-time performance in such multi-core and many-core systems, a difficult problem is to analyze the execution flows of tasks that may access both CPU cores and shared resources. A common approach is to analyze schedulability from a conventional core-centric view [4] [5] [6]. That is, the tasks are scheduled on CPU cores under certain scheduling algorithms, which integrate the latency due to resource contention in the analysis. The intuitive idea is to focus on judiciously allocating CPU resources while viewing shared resources like I/O and bounding the worst-case latency a task may experience on such resources. Then, by treating such latency as suspension delays, we can transform the original multi-resource scheduling problem into single-resource (i.e., CPU) scheduling of suspending tasks.

For a more in-depth treatment of the literature on the impact of resource sharing on performance and worst-case timing analysis, a survey can be found in [7]. One line of work in timing analysis for multicore systems is based on an assumption that the CPU execution and shared-resource access patterns are well structured, e.g., [8] [9] [10] [11] [12]. For example, in the superblock execution model [11], the execution of a superblock has three phases: data acquisition, local execution, and data replication phases. A similar model can be found in the open international standard IEC 61131-3. Altmeyer et. al [13] present a framework to decouple response-time analysis in a compositional manner based on the context-independent WCET values.

The CPU-centric perspective is sound for traditional embedded systems where computing resources may be insufficient while the contention on shared resources is often light. However, for advanced embedded systems that employ multi-core platforms to serve large-scale real-time workloads, shared

resources may become the actual scheduling bottleneck, causing the worst-case latency bound on the shared resource (thus the resulting schedulability test) to be rather pessimistic or even impossible to derive. Motivated by this observation, it may be much more viable to resolve this multi-resource scheduling problem from the counter-intuitive shared-resource centric perspective since tasks may experience much lighter contention on CPU cores, as previously argued in [14] which considers a simplified single-unit shared resource scenario (equivalent to uniprocessor scheduling from a shared-resource-centric perspective).

To resolve the multi-resource scheduling problem, researchers start investigating it from the above-mentioned shared-resource-centric perspective. This means novel techniques will be developed to focus on judiciously scheduling tasks' requests on the shared resource (treating the shared resource as the first-class units) and bounding the worst-case latency a task may experience on the CPU cores (treating CPU cores as "I/O"). The benefit is obvious, particularly on many-core platforms, where tasks may receive rather trivial or even no interference on CPU cores as the CPU resource is often abundant compared to the shared resource. Researchers seek to develop practical solutions that may benefit embedded systems designers. A recent work [15] has developed a comprehensive set of techniques that are based on non-trivial schedulability tests with different runtime complexity and accuracy (w.r.t. both schedulability and the needed size of the shared resource) for scheduling a set of hard real-time (HRT) tasks that may access CPU cores and multiple units of a shared resource in an interleaving manner. If reasonably high runtime/analysis complexity is affordable, then their solutions may yield a high schedulability and a minimized required size of the shared resource, which is one of the most critical factors that may reduce the overall cost and complexity of the SWaP (size, weight, and power) constrained embedded systems.

- [1] C. Clauss, S. Lankes, P. Reble, and T. Bemberl. Evaluation and improvements of programming models for the intel scc many-core processor. In High Performance Computing and Simulation (HPCS), 2011 International Conference on, pages 525–532. IEEE, 2011
- [2] D.-R. Fan, N. Yuan, J.-C. Zhang, Y.-B. Zhou, W. Lin, F.-L. Song, X.-C. Ye, H. Huang, L. Yu, G.-P. Long, et al. Godson-t: An efficient manycore architecture for parallel program executions. *Journal of Computer Science and Technology*, 24(6):1061, 2009.
- [3] F. Thabet, Y. Lhuillier, C. Andriamisaina, J.-M. Philippe, and R. David. An efficient and flexible hardware support for accelerating synchronization operations on the sthorm many-core architecture. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 531–534. EDA Consortium, 2013
- [4] A. Biondi, A. Balsini, M. Pagani, E. Rossi, M. Marinoni, and G. Buttazzo. A framework for supporting real-time applications on dynamic reconfigurable fpgas. In *Real-Time Systems Symposium (RTSS)*, 2016 IEEE, pages 1–12. IEEE, 2016.
- [5] Z. Dong and C. Liu. Closing the loop for the selective conversion approach: A utilization-based test for hard real-time suspending task systems. In *Real-Time Systems Symposium (RTSS)*, 2016 IEEE, pages 339–350. IEEE, 2016.
- [6] R. Pellizzoni and H. Yun. Memory servers for multicore systems. In *Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2016 IEEE, pages 1–12. IEEE, 2016.
- [7] A. Abel, F. Benz, J. Doerfert, B. Dorr, S. Hahn, F. Hauptenthal, M. Jacobs, A. H. Moin, J. Reineke, B. Schommer, and R. Wilhelm. Impact of resource sharing on performance and performance prediction: A survey. In P. R. D' Argenio and H. Melgratti, editors, *CONCUR 2013– Concurrency Theory*, volume 8052 of *Lecture Notes in Computer Science*, pages 25–43. 2013
- [8] N. Guan, P. Ekberg, M. Stigge, and W. Yi. Resource sharing protocols for real-time task graph systems. In *Real-Time Systems (ECRTS)*, 2011 23rd Euromicro Conference on, pages 272–281. IEEE, 2011
- [9] J.-J. Han, D. Zhu, X. Wu, L. T. Yang, and H. Jin. Multiprocessor realtime systems with shared resources: Utilization bound and mapping. *IEEE Transactions on Parallel and Distributed Systems*, 25(11):2981–2991, 2014
- [10] W.-H. Huang, J.-J. Chen, H. Zhou, and C. Liu. Pass: Priority assignment of real-time tasks with dynamic suspending behavior under fixed-priority scheduling. In *Proceedings of the 52nd Annual Design Automation Conference*, page 154. ACM, 2015.

- [11] R. Pellizzoni, A. Schranzhofer, J.-J. Chen, M. Caccamo, and L. Thiele. Worst case delay analysis for memory interference in multicore systems. In DATE, pages 741–746, March 2010.
- [12] A. Schranzhofer, R. Pellizzoni, J.-J. Chen, L. Thiele, and M. Caccamo. Timing analysis for resource access interference on adaptive resource arbiters. In RTAS, pages 213–222, April 2011.
- [13] S. Altmeyer, R. I. Davis, L. S. Indrusiak, C. Maiza, V. Nelis, and J. Reineke. A generic and compositional framework for multicore response time analysis. In Proceedings of the 23rd International Conference on Real Time Networks and Systems, RTNS 2015, Lille, France, November 4-6, 2015, pages 129–138, 2015.
- [14] W.-H. Huang, J.-J. Chen, and J. Reineke. MIRROR: symmetric timing analysis for real-time tasks on multicore platforms with shared resources. In Proceedings of the 53rd Annual Design Automation Conference, DAC 2016, Austin, TX, USA, June 5-9, 2016, pages 158:1– 158:6, 2016.
- [15] Zheng Dong, Cong Liu, Soroush Bateni, Kuan-Hsun Chen, Jian-Jia Chen, Georg von der Brüggen, and Junjie Shi. "Shared-resource-centric limited preemptive scheduling: A comprehensive study of suspension-based partitioning approaches." In 2018 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), pp. 164-176. IEEE, 2018.

[Back to Contents](#)

Paper Submission Deadlines

ISQED'21 - Int'l Symposium on Quality Electronic Design
Santa Clara, CA
Deadline: Oct 2, 2020
Apr , 2021
<http://www.isqed.org>

SLIP^2 - System-Level Interconnect Problems and Pathfinding (co-located with ICCAD 2020)
San Diego, CA
Deadline: Oct 3, 2020
Nov 5, 2020
<http://sliponline.org>

ISPD' 21 – ACM Int' l Symposium on Physical Design (canceled)
Deadline: Oct 12, 2020 (Abstracts due: Oct 2, 2020)
Mar 21-24, 2021
<http://www.ispd.cc>

RTAS'21 – 27th IEEE Real-Time and Embedded Technology and Applications Symposium
Nashville, USA
Deadline: Oct 26, 2020 (Firm)
May 18-21, 2021
<http://2021.rtas.org>

RTAS is a top-tier conference in real-time systems listed in csranking. It focuses on systems research related to embedded systems and time-sensitive systems (of any size). The scope of RTAS ranges from traditional hard real-time systems to embedded systems without explicit timing requirements, including latency-sensitive systems with informal or soft real-time requirements.

ISCA' 21 – Int' l Symposium on Computer Architecture
Valencia, Spain
Deadline: Nov 24, 2020 (Abstracts due: Nov 19, 2020)
May 22 – 26, 2021
<https://iscaconf.org/isca2021/>

ISED' 21 – 10th Int' l Symposium on Embedded Computing & System Design
Kollam, India

Deadline: Jan 5, 2021

Mar 12-14, 2021

<http://isedconf.org>

TAU' 21 – ACM Int' l Workshop on Timing Issues in the Specification and Synthesis of Digital Systems

Monterey, CA

Deadline: Jan 9, 2021

Apr 8-9, 2021

<http://www.tauworkshop.com>

FCCM' 21 - The 29th IEEE International Symposium On Field-Programmable Custom Computing Machines

Orlando, FL

Deadline: Jan 11, 2021 (Abstracts due: Jan 4, 2021)

May 9 – May 12, 2021

<https://www.fccm.org/>

[Back to Contents](#)

Upcoming Conferences and Symposia

PACT'20 - Int'l Conference on Parallel Architectures and Compilation Techniques

Virtual Conference

Oct 2, 5-7, 2020

<http://www.pactconf.org>

VLSI-SoC' 20 – IFIP/IEEE Int' l Conference on Very Large Scale Integration

Virtual Conference

Oct 5-9, 2020

<http://www.vlsi-soc.com>

ISCAS'20 – IEEE Int'l Symposium on Circuits and Systems

Virtual Conference

Oct 10-21, 2020

<http://iscas2020.org>

MICRO' 20 – IEEE/ACM Int'l Symposium on Microarchitecture

Virtual Conference

Oct 17-21, 2020

<http://www.microarch.org/micro53>

BodyNets'20 – Int' l Conference on Body Area Networks

Virtual Conference

Oct 21-22, 2020

<http://www.bodynets.org>

ICCAD' 20 – IEEE/ACM Int' l Conference on Computer-Aided Design

Virtual Conference

Nov 2-5, 2020

<http://www.iccad.com>

WOSET'20 - Workshop on Open-Source EDA Technology (co-located with ICCAD 2020)

San Diego, CA

Nov 5, 2020

<https://woset-workshop.github.io>

HOST'20 – IEEE Int'l Symposium on Hardware-Oriented Security and Trust
Virtual Conference

Dec 7-11, 2020

<http://www.hostsymposium.org>

FPT'20 - Int'l Conference on Field-Programmable Technology
Virtual

Dec 7-11, 2020

<http://icfpt.org>

HiPC'20 – IEEE Int'l Conference on High Performance Computing, Data, And Analytics
Pune, India

Dec 16-19, 2020

<http://www.hipc.org>

iSES' 20 – IEEE Int'l Symposium on Smart Electronic Systems
Chennai, India

Dec 14-16, 2020

<http://www.ieee-ises.org>

ASP-DAC'21 - Asia and South Pacific Design Automation Conference
Virtual Conference

Jan 18-21, 2021

<http://www.aspdac.com>

HiPEAC'21: Int'l Conference on High Performance Embedded Architectures & Compilers
Budapest, Hungary

Jan 18-20, 2021

<https://www.hipeac.net/2021/budapest>

DATE'21 - Design Automation and Test in Europe
Grenoble, France

Feb 1-5, 2021

<http://www.date-conference.com>

ISSCC'21 – IEEE Int'l Solid-State Circuits Conference
San Francisco, CA

Feb 14-18, 2021

<http://isscc.org>

VLSID'21 – International Conference on VLSI Design & International Conference on Embedded
Systems

Virtual Conference

Feb 20-24, 2021

<http://embeddedandvlsidesignconference.org/>

FPGA' 21 – ACM/SIGDA Int'l Symposium on Field-Programmable Gate Arrays
Virtual Conference

Feb 28-Mar 2, 2021

<http://www.isfpga.org>

[Back to Contents](#)

Researcher Spotlight

Hello readers,

Welcome to Researcher spotlight. In this edition we meet Prof. Iris Bahar, Professor of Computer Science with the Department of Computer Science and a Professor of Engineering with the School of Engineering both at Brown University, Providence, Rhode Island. Prof. Bahar leads the Laboratory for Engineering Man/Machine Systems (LEMS). She received B.S. in Computer Engineering and M.S. in Electrical Engineering both from the University of Illinois, Urbana and Ph.D. in Electrical Engineering from the University of Colorado, Boulder, Colorado. Below are excerpts from a recent conversation.

1. Can you share with us some of the research areas you are interested in?

My research interests lie broadly in the areas of computer system design and electronic design automation. In particular, my research focuses on energy-efficient and reliable computing, from the system level to device level. Past research topics have included modelling thermal noise effects in nanoscale circuits, design of noise- and error-immune circuits, approximate computing (from systems to circuits), and memory synchronization techniques for multiprocessor systems. Most recently, my research interests have led me to explore applications for near-data processing and design of robust machine learning techniques for robot scene perception.

2. Verification of logic through testbenches by itself requires exhaustive cycles of simulation. Design variants immune to noisy compute may require a customization of verification intent as well. Managing logical bugs vs induced bugs vs maintaining reasonable verification complexity - collectively these seem nearly conflicting goals. How do you determine an optimality amongst these?

Scaling of semiconductor devices has enabled higher levels of integration and performance improvements at the price of making devices more susceptible to the effects of static and dynamic variability. Adding safety margins (guardbands) on the operating frequency or supply voltage prevents timing errors but has a negative impact on performance and energy consumption. As a means of managing this pessimism, we proposed a novel HW/SW technique that relies on Hardware Transactional Memory rollback mechanisms for error correction in errant transactions. Hardware Transactional Memory (HTM) was originally proposed for managing memory synchronization in multiprocessor systems by providing a means to speculate about shared data protection to improve program runtime performance. Our approach replaces traditional conflict detection logic with simpler architectural support for error detection and employs error management policies that aggressively apply dynamic voltage scaling (DVS) beyond the point of first failure for better energy savings. The policy monitors transaction aborts and commits to estimate the experienced error rate and decides whether to lower, maintain or raise the voltage level.

With this proposed technique comes the extension of using the same HTM framework to manage approximate computing. Approximate computing has emerged as a promising solution to these dilemmas for applications that can sustain a slightly reduced accuracy for increases in performance and energy efficiency; however, managing this approximation dynamically within an application can be a challenge of its own. If not done correctly, approximations may lead to unacceptable quality loss, or worse, it can affect critical data and damage the control flow of the program. Our same HTM-inspired framework provides a novel error management scheme that tolerates (i.e., opportunistically ignores) timing violations, allowing for more aggressive voltage scaling. Dynamically deciding which timing violations to ignore relies on careful evaluation of the application running of the system as well as developing an accurate error model to capture the error behavior within the processor computation flow. In our work, our error model takes into account value correlation, computation history, and the critical path of the computation to more accurately determine if a particular error in space and time is critical or not. We then utilize a combination of static and dynamic monitors to determine appropriate conditions for voltage adjustments within tolerable bounds based on this

error analysis . The key insight is that recovery from critical errors, ones that cannot be tolerated, can be facilitated by lightweight mechanisms adapted from hardware transactional memory (HTM) to optimize energy savings while retaining similar runtime performance at acceptable accuracy loss. Our experimental results show our approach allows up to 47% total energy savings with negligible impact on runtime.

We note that our approach requires special circuitry to detect timing errors during program execution and our error models use assumptions in the hardware design to evaluate the impact of errors on application accuracy. Once the models are determined, the effect of errors on the accuracy of a particular application requires profiling the application offline to characterize which instructions are amenable to approximation and the extent to which the application can tolerate errors. Of course, this puts some extra burden on the user for understanding what can be approximated and what testbenches are appropriately representative of actual application use. On the other hand, the hardware infrastructure we propose does not change with the error model or the results of the error analysis; any updates to error models or testbenches will not change in our hardware. This is distinct from other works that propose special approximate hardware circuits as part of the design.

3. Computing nodal voltages using differential equations and solving them seems an efficient way to analyse thermal noise transients. The speedup against projected SPICE simulations is also very good. How does this approach fare in terms of accuracy?

Near-threshold and sub-threshold voltage designs have been identified as possible solutions to overcome the limitations introduced by energy consumption in modern VLSI circuits. However, aggressive voltage and gate length scaling will reduce the reliability of logic circuits due to the increasing impact of noise and variability effects. Therefore, designers need new tools to simulate logic circuits in the presence of noise. Time-domain analysis helps understand how transient faults affect a circuit and can guide designers in producing noise-resistant circuitry. However, standard approaches such as SPICE that can be used to model intrinsic noise sources in the time domain are computationally expensive. Moreover, small noise-driven fluctuations in electron occupation of circuit nodes introduce time-varying biasing point fluctuations, increasing the modeling complexity. To address these challenges, we have proposed a new approach to modeling thermal noise and random telegraph signal (RTS) noise directly in the time domain by developing and solving a series of stochastic differential equations (SDEs). In comparisons to traditional SPICE-based simulations, our approach can provide 3 orders of magnitude speedup in simulation time without sacrificing accuracy. Moreover, we have also introduced a novel, iterative threshold-crossing algorithm, aimed at the efficient sampling of rare noise transients. We have shown that Monte-Carlo simulations based on this approach can detect rare high-amplitude single event transients (SETs) that would be impossible to uncover with standard SPICE-based transient simulators.

Finally, we have extended our approach to analyze the reliability of latches and SRAMs operating in subthreshold conditions. We were able to evaluate how reliability due to thermal noise of SRAMs built from 7nm FinFET technology were affected by lowered supply voltage, increased process variability, and temperature shifts. Our work has made it possible to quantitatively evaluate in minutes the asymptotic behavior of extremely rare error events that lead to bit-flip errors. Again, such an analysis would be impossible with conventional SPICE-based transient simulation methods. Another nice feature of our work is that our simulation framework can be adapted to other technology nodes or other sources of noise and provides a means of evaluating the trade-offs in the robust design of low-power SRAM for ultra-low-power memory.

4. The idea of bringing memory and compute closer to each other has an analogy in the FPGA world in the form of apportioning silicon area between Look Up Table and Configuration Memory. In the case of compute logic, how do you envision apportioning. In other words, how much logic is too much for a given unit of memory (or vice versa)?

Recent advances in memory architectures have provoked renewed interest in near-data-processing (NDP) as way to alleviate the “memory wall” problem. An NDP architecture places logic circuits,

such as simple processors, in close proximity to memory. This is distinct from processing-in-memory (PIM) where logic computation is effectively integrated into the memory cells/arrays. PIM architectures are analogous to FPGA fabrics where a Configurable Logic Block composed of a lookup table (for combinational logic) and flip-flops (for sequential logic), but with much tighter integration between logic and memory. 3D die-stacking technology allows logic circuits, such as simple processors, to be placed physically near memory, using high-bandwidth Through-Silicon Via (TSV) interconnects for communication between the near-memory processor and memory. Today, commercially available devices that exploit the 3D die-stacking technology, such as Hybrid Memory Cube (HMC) or High-Bandwidth Memory (HBM), implement only simple memory controller logic near memory. However, we expect that soon simple processors will be placed near memory, enabling near-data-processing. We have investigated how near-memory accelerators can be combined with novel data structures and algorithms to exploit the low-latency, high-bandwidth memory access of future NDP architectures, while also preserving the high concurrency of conventional systems. In particular, we have focused on software libraries and architectural support for general-purpose concurrent data structures with near-data-processing architectures. These data structures are used in many applications and adapting them to NDP architectures is a key step toward making these architectures useful. In conventional architectures, “pointer-chasing” data structures with poor cache locality and high-contention concurrent data structures are often bottlenecks, while near-data-processing architectures have the promise to alleviate or even eliminate these problems. We found that potential benefits of NDP-based concurrent data structures also required lightweight NDP hardware modifications (inspired by observations on data structure access patterns and underlying DRAM activity). Our software-hardware approach showed significant improvements in performance and energy consumption compared to state-of-the-art concurrent data structures.

5. The idea of combining convolutional neural networks with sampling- and probabilistic-based techniques seems to provide a real advantage in terms of improved accuracy. Why can't the same accuracy improvements be attained with better or more training of the neural network?

Technological advancements have led to a proliferation of robots using machine learning systems to assist humans in a wide range of tasks. However, we are still far from accurate, reliable, and resource-efficient operations of these systems. Despite the strengths of convolutional neural networks (CNNs) for object recognition, these discriminative techniques have several shortcomings that leave them vulnerable to exploitation from adversaries, such as their need for extremely large training sets, their “black box” decision making, and their inability to recover from incorrect inferences. Moreover, CNNs tend to overfit to the training data due to their high non-linearity and parameter counts. Overfitting also makes the CNN vulnerable to adversarial attack (e.g., via small image perturbations), and can also lead to poor predictions when faced with unfamiliar scenarios. In particular, when the robot operates in the real world, it is subject to complex and changing environments that often have not been captured by training data. Finally, the computational, financial, and environmental cost incurred to train these discriminative models can be quite immense, as they often require weeks or even months to adequately train.

In contrast, generative probabilistic inference techniques such as Monte-Carlo sampling are inherently explainable, general, and resilient through the process of generating, evaluating, and maintaining a distribution of many hypotheses representing possible decisions. Unfortunately, this robustness comes at the cost of computational efficiency. Alternatively, our work using hybrid discriminative-generative approaches offers a promising avenue for robust perception and action. Such methods combine inference by deep learning with sampling and probabilistic inference models, and the ability to represent actual and counterfactual experiments to achieve robust and adaptive understanding. This hybrid approach allows intelligent systems to reason about, interact with, and manipulate objects in complex (and even adversarial) environments. Our experiments have shown that our approach can achieve up to 40% improvement in pose estimation accuracy compared to end-to-end neural network approaches, demonstrating robust performance especially in dark or occluded environments.

While neural network inference can be completed within a second on modern general-purpose

graphic processing units (GPUs), the iterative process of Monte-Carlo sampling does not map well to GPU acceleration, making the algorithm less amenable to meeting the energy and real-time constraints required of mobile applications. In particular, the run time and energy consumption is determined by the range of sampling, the number of iterations, and the computational complexity of the likelihood function. To address this challenge, we have developed novel hardware accelerated implementations of Monte-Carlo sampling on FPGA fabrics. The main benefits of using an FPGA is the configuration of all the resources near one other on the same fabric and pipelining the data processing across the various steps of the algorithm. This results in less item spent for data transfers, less need to store intermediate results between steps, and more opportunity for parallel execution. With our FPGA implementation, we can achieve real-time performance without sacrificing accuracy and with significantly reduced energy consumption. In particular, our design runs 30% faster than a high-end GPU implementation with only 2% of the energy consumption, and 95% faster than a low-power GPU implementation, dissipating approximately the same amount of power but with only 4% of the energy consumption. We find this work very promising and plan to continue investigating hardware acceleration of other generative algorithms to be combined with discriminative (i.e., neural network) techniques.

[Back to Contents](#)

SIGDA Partner Journal

1. TODAES Best Paper Award and Trending Articles

The ACM Transactions on Design Automation of Electronic Systems (TODAES), the premier ACM journal in design and automation of electronic systems and a closer partner of SIGDA, has announced its Best Paper Award. Congratulations to Bo-Yuan Huang (Princeton University, New Jersey, USA), Hongce Zhang (Princeton), Pramod Subramanyan (Indian Institute of Technology Kanpur, India), Yakir Vizel (Technion Israel Institute of Technology, Haifa, Israel), Aarti Gupta (Princeton), and Sharad Malik (Princeton) for their article "Instruction-Level Abstraction (ILA): A Uniform Specification for System-on-Chip (SoC) Verification". In this work, they address challenges in verification of "accelerator-rich" heterogeneous System on Chip (SoC) platforms by providing a formal specification and high-level abstraction for accelerator functional behavior. Their formalized Instruction Level Abstraction framework permits equivalence checking between two ILAs, as well as an ILA and its finite state machine implementation, and supports accelerator upgrades.

TOADAES also released its 6th issue in 2020 (Volume 25). In this newsletter, we will discuss two articles that have been highly downloaded by TODAES readers since the issue's release. You are invited to visit the TODAES homepage at <https://dl.acm.org/journal/todaes>, where you may read these articles, as well as other innovative work documenting significant research and development advances in electronic system design, emphasizing a computer science/engineering orientation. You are also invited to submit research in these areas, and theoretical analysis and practical solutions are welcome.

In the article, "Energy-Efficient GPU L2 Cache Design Using Instruction-Level Data Locality Similarity", authors Jingweijia Tan, Kaige Yan (Jilin University, Jilin, China), Shuaiwenleon Song (Pacific Northwest National Laboratory, Richland, WA, USA), and Xin Fu (University of Houston, Houston, TX, USA) present an energy-efficient cache design for massively parallel, throughput-oriented architectures. They identified that 95.6% of data stored in L2 cache is under utilized. They applied SIMT programming to achieve instruction-level data locality similarity, which can be used to accurately predict the data re-reference counts at L2 cache block level. Their experimental results of their design approach, called LOSCache, indicate a significant reduction of L2 cache energy dissipation by an average of 64% with a mere 0.5% performance tradeoff.

Subodha Charles and [Prabhat Mishra](#) (University of Florida, Gainesville, FL, USA) presented their work "Reconfigurable Network-on-Chip Security Architecture", which focuses on complex SOC's being used in edge devices in IoT applications. Most current approaches to securing Network-on-Chip (NoC)

devices, such as lightweight encryption and authentication, are statically optimized security solutions. Dynamic requirement changes and long IoT application life make these security approaches difficult to implement in the real world. They analyzed current architecture and threat models, and present a tiered system of security primitives and corresponding reconfigurable parameters depending on use-case scenarios. Then, they designed an efficient reconfiguration architecture that uses security agents to monitor system environment characteristics to decide which security mechanisms to activate based on security policies.

2. ACM TRETs Special Issue on FPGAs in data centers

Guest Editors

Ken Eguro, Microsoft

Stephen Neuendorffer, Xilinx

Viktor Prasanna, University of Southern California

Hongbo Rong, Intel

Hardware accelerators have been used recently to augment the computing power of data centers to improve performance of many applications, particularly to optimize latency sensitive applications. In fact, several commercial vendors offer FPGAs in their cloud platforms.

In this special section of TRETs, we call for advanced research in using FPGAs in data centers. Topics of interest include (but are not limited to) the following:

- Programming languages, Compilers, Libraries, Runtime scheduling systems, Tools, etc. for targeting FPGAs for application acceleration in data centers
- Large-scale problem solving using many FPGAs in data centers, e.g. linear algebra with huge matrices, data bases, serverless computing, large-scale machine learning, etc.
- Virtual Overlays and implementations of applications on FPGAs in data centers, e.g. retargetable high-performance overlays for heterogeneous FPGAs of different vendors or models
- Latency and performance tradeoffs in using FPGAs for acceleration
- FPGA IP cores for data center acceleration
- Novel applications of FPGAs in data centers
- Virtualization and run-time resource management in using FPGAs
- End to end application acceleration on FPGAs in data centers
- Comparison studies of using FPGAs in data centers using other acceleration architectures (GPUs, TPUs, ASICs, CPUs, etc.)
- Communication optimization using FPGAs in conjunction with other computing platforms, e.g. Smart NICs, Software Defined Networking (SDN), etc.
- Storage optimization using FPGAs
- Use and management of network accessible FPGAs
- Survey and tutorial studies of FPGAs in data centers

Submission Deadline: December 1, 2020

Target Publication Time: Summer, 2021

For more information on this special section, please contact prasanna@usc.edu.

Manuscripts are to be submitted at mc.manuscriptcentral.com/trets. Please select the paper type "Special Issue on FPGAs in Data Centers" when submitting your paper.

[Back to Contents](#)

Technical Activities

1. "6 Accelerating Tech Trends Due to Covid-19"

The Covid-19 pandemic continues to have a profound impact on our everyday lives, on most industries, and the economy in general...

[\[https://www.eetimes.eu/6-accelerating-tech-trends-due-to-covid-19/\]](https://www.eetimes.eu/6-accelerating-tech-trends-due-to-covid-19/)

2. "On Demand Webinar: How FPGAs Enable Industrial Automation and Transformation to Industry 4.0"

In a recently published webinar, the folks at Intel discussed current trends in Industry 4.0 and the key technologies enabling the growth of this new market...

[\[https://iot.eetimes.com/on-demand-webinar-how-fpgas-enable-industrial-automation...\]](https://iot.eetimes.com/on-demand-webinar-how-fpgas-enable-industrial-automation...)

3. "The Era Of Big Memory Is Upon US"

If you reduce systems down to their bare essentials, everything exists in those systems to manipulate data in memory, and like human beings, all that really exists for any of us is what is in memory...

[\[https://www.nextplatform.com/2020/09/23/the-era-of-big-memory-is-upon-us/\]](https://www.nextplatform.com/2020/09/23/the-era-of-big-memory-is-upon-us/)

4. "AI-Based Sequence Detection for IP and SoC Verification and Validation"

A couple of years ago at the Design Automation Conference (DAC), as I walked the exhibit floor I was amused by how many EDA vendors had jumped on the marketing bandwagon for artificial intelligence (AI) and machine learning (ML)...[\[https://www10.edacafe.com/blogs/agnisys/\]](https://www10.edacafe.com/blogs/agnisys/)

5. " 'Legacy' Memories Ready to Drive Disruptive Innovation"

Legacy memories are no longer lowly devices that hit their end of life (EOL) because a major vendor is focused on the latest and greatest...

[\[https://www.eetasia.com/legacy-memories-ready-to-drive-disruptive-innovation/\]](https://www.eetasia.com/legacy-memories-ready-to-drive-disruptive-innovation/)

Job Openings:

1. University of California Los Angeles

Job Title: Adjunct Professor in Computer Science

Description: The Department of Computer Science in the Henry Samueli School of Engineering and Applied Science at University of California, Los Angeles, invites applications for part-time positions as an adjunct professor. These areas include, but are not limited to, Artificial Intelligence, Computer System Architecture & CAD, Computational Systems Biology, Graphics and Vision, Information & Data Management, etc. A Ph.D. in Computer Science or an equivalent degree is required. The University of California, Los Angeles is an Equal Opportunity/Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, national origin, disability, age or protected veteran status. For the complete University of California nondiscrimination and affirmative action policy, see: UC Nondiscrimination & Affirmative Action Policy at: <http://policy.ucop.edu/doc/4000376/NondiscrimAffirmAct>. Learn More More information about this recruitment: <https://www.cs.ucla.edu/>

2. Indiana University Bloomington United States

Job Title: Assistant Professor in Computer Science

Description: The Luddy School of Informatics, Computing, and Engineering at Indiana University (IU) Bloomington invites applications for a tenure track assistant professor position in Computer Science to begin in Fall 2021. We are particularly interested in candidates with research interests in formal models of computation, algorithms, information theory, and machine learning with connection to quantum computation, quantum simulation, or quantum information science. The successful candidate will also be a Quantum Computing and Information Science Faculty Fellow supported in

part for the first three years by an NSF-funded program that aims to grow academic research capacity in the computing and information science fields to support advances in quantum computing and/or communication over the long term. Applicants should have a demonstrable potential for excellence in research and teaching and a PhD in Computer Science or a related field expected before August 2021. Questions may be sent to sabry@indiana.edu

3. University of Waterloo Canada

Job Title: Associate/Full Professor of Experimental Quantum Information Science or Engineering

Description: IQC is a collaborative research institute at the University of Waterloo focused on quantum information science and technology, ranging from the theory of quantum information to practical applications. At present, IQC has a complement of 32 faculty members 65 postdoctoral fellows and 150 graduate students from the Faculties of Engineering, Mathematics and Science. Membership in IQC is renewable and comes with research space, a teaching reduction of one course per year, and a stipend. Information about research at IQC can be found at uwaterloo.ca/iqc/research, and tqt.uwaterloo.ca. Waterloo is extraordinarily rich in quantum activities, so much so that it has been branded the "Quantum Valley." For IQC faculty, there are opportunities to engage with the Perimeter Institute for Theoretical Physics, the Quantum Valley Ideas Lab, and the many young quantum companies located around the university. At Waterloo, you will have the opportunity to work across disciplines and collaborate with an international community of scholars and a diverse student body, situated in a rapidly growing community that has been termed a "hub of innovation." All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority. If you have any questions regarding the position, the application process, assessment process, eligibility, or a request for accommodation during the hiring process, please contact the IQC Director at iqc-dtr@uwaterloo.ca. Three reasons to apply: uwaterloo.ca/faculty-association/why-waterloo.

4. Guangdong Technion Israel Institute of Technology China

Job Title: Assistant/Associate /Full Professorships of Engineering

Description: Israel Institute of Technology manages the academics of a new university established in Shantou City, Guangdong Province, China gtiit.edu.cn/en. GTIIT follows the academic model of the Technion campus in Haifa, Israel. It is dedicated to high-quality research and education in science and technology. Courses will be taught in English. GTIIT has established undergraduate programs in Chemical Engineering, Biotechnology & Food Engineering and Materials Engineering. An undergraduate program in Mathematics (major) with Computer science (minor) is expected to start in October 2020. Guangdong Technion branch intends to launch an undergraduate education program in Mechanical Engineering, which will include Tracks for Robotics, Mechanics of Materials and Energy. A graduate program in these areas will also be offered in coordination with the Technion's Graduate School. GTIIT is also opening/establishing the Centers for Robotics, Science & Engineering in Health and Medicine and Sustainable World. Its members are expected to develop vigorous research program and commit to high-level teaching. Tenure-track faculty positions at all ranks are now available for exceptional candidates. Please send CV, list of publications, research plan and teaching statements to the Search Committee at gtiitrecruit@technion.ac.il

[Back to Contents](#)

Notice to Authors

Notice to Authors

By submitting your article for distribution in this Special Interest Group publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights: to publish in print on condition of acceptance by the editor; to digitize and post your article in the electronic version of this publication; to include the article in the ACM Digital Library and in any Digital Library related services; and to allow users to make a personal copy of the article for noncommercial, educational or research purposes. However, as a contributing author, you retain copyright to your article and ACM will refer requests for republication directly to you.

This newsletter is a free service for current SIGDA members and is added automatically with a new SIGDA membership.

Circulation: 2,700

This ACM/SIGDA E-NEWSLETTER is being sent to all persons on the ACM/SIGDA mailing list. To unsubscribe, send an email to listserv@listserv.acm.org with "signoff sigda-announce" (no quotes) in the body of the message. Please make sure to send your request from the same email as the one by which you are subscribed to the list.